Differentiable Probabilistic Models

William Watson
nextbillyonair@gmail.com

Abstract

d

Contents

1 Introduction **7**

1.1 Philosophy 7

2 Preliminary **7**

2.1 Kronecker Product 7

2.2 Gradients 7

2.3 Jacobian 7

2.4 Hessian 7

2.5 Newton Optimization 7

3 Distributions **7**

3.1 Distribution 7

3.2 Arcsine 7

3.3 Asymmetric Laplace 7

3.4 Bernoulli 7

3.5 Beta 7

3.6 Categorical 7

3.7 Cauchy 7

3.8 Chi Square 7

3.9 Conditional Model 7

3.10 Convolution 7

3.11 Data 7

3.12 Dirac Delta 7

3.13 Dirichlet 7

3.14 Exponential 7

3.15 Fisher-Snedcor (F-Distribution) 7

3.16 Gamma 7

3.17 Generator 7

3.18 Gumbel Softmax 7

3.19 Gumbel 7

3.20 Half Cauchy 7

3.21 Half Normal 7

3.22 Hyperbolic Secant 7

3.23 Langevin 7

3.24 Laplace 7

3.25 Log Cauchy 7

3.26 Log Laplace 7

3.27 Log Normal 7

3.28 Logistic 7

3.29 Logit Normal . . . 7
3.30 Normal (Multivariate) . . . 7
3.31 Rayleigh . . . 7
3.32 Relaxed Bernoulli . . . 7
3.33 Student T . . . 7
3.34 Transformed Distribution . . . 7
3.35 Uniform . . . 7

4 Mixture Models **7**
4.1 Mixture Model . . . 8
4.2 Gumbel Mixture Model . . . 8
4.3 Infinite Mixture Model . . . 8

5 Transforms **8**
5.1 Transform . . . 8
5.2 Inverse Transform . . . 8
5.3 Affine . . . 8
5.4 Exp . . . 8
5.5 Expm1 . . . 8
5.6 Gumbel . . . 9
5.7 Log . . . 9
5.8 Logit . . . 9
5.9 Planar . . . 9
5.10 Power . . . 9
5.11 Reciprocal . . . 10
5.12 Sigmoid . . . 10
5.13 SinhArcsinh . . . 10
5.14 Softplus . . . 10
5.15 Softsign . . . 10
5.16 Square . . . 10
5.17 Tanh . . . 10

6 Criterion and Divergences **10**
6.1 Cross-Entropy . . . 10
6.2 Perplexity . . . 10
6.3 Forward KL Divergence . . . 11
6.4 Reverse KL Divergence . . . 11
6.5 Jensen-Shannon Divergence . . . 11
6.6 Earth Mover's Distance . . . 11

7 ELBO **11**

8 Adversarial Loss **11**
8.1 Adversarial Loss . . . 13
8.2 GAN Loss . . . 13
8.3 MMGAN Loss . . . 13
8.4 WGAN Loss . . . 13
8.5 LSGAN Loss . . . 13
8.6 Gradient Penalty . . . 13
8.7 Spectral Norm . . . 13

9 Models **13**
9.1 Base Models . . . 13
9.1.1 Model . . . 13
9.2 Regression . . . 13
9.2.1 Linear Regression (Normal) . . . 13
9.2.2 L1 Regression (Laplace) . . . 13
9.2.3 Ridge Regression (Normal + Normal Prior on Weights) . . . 13
9.2.4 Lasso Regression (Normal + Laplace Prior on Weights) . . . 13
9.3 Classification . . . 13
9.3.1 Logistic Regression (Bernoulli) . . . 13
9.3.2 Bayesian Logistic Regression (Bernoulli) . . . 13
9.3.3 Softmax Regression (Categorical) . . . 13
9.3.4 Bernoulli Naive Bayes (Bernoulli - Bernoulli) . . . 13
9.3.5 Gaussian Naive Bayes (Multinomial - Gaussian) . . . 13
9.3.6 Multinomial Naive Bayes (Bernoulli - Multinomial) . . . 13
9.3.7 Linear Discriminant Analysis (Multinomial - Shared Covariance) . . . 13
9.3.8 Quadratic Discriminant Analysis (Multinomial - Multivariate Gaussian) . . . 13
9.4 Clustering . . . 13
9.4.1 Gaussian Mixture Model . . . 13
9.5 Unconstrained Matrix Factorization (Gaussian) . . . 13
9.6 Principle Components Analysis . . . 13
9.6.1 PCA . . . 13
9.6.2 EM-PPCA . . . 13
9.6.3 Variational PPCA . . . 13
9.7 Generative Adversarial Networks . . . 13
9.7.1 Generative Adversarial Networks . . . 13
9.7.2 GAN Model . . . 13
9.7.3 MMGAN Model . . . 13
9.7.4 WGAN Model . . . 13
9.7.5 LSGAN Model . . . 13
9.8 Variational Auto-Encoders (TBD) . . . 13

10 Monte Carlo **13**

10.1 Monte Carlo Approximation 13

10.2 Linear Congruential Generator 13

10.3 Inverse Transform Sampling 13

10.4 Box-Muller 13

10.5 Marsaglia-Bray 13

10.6 Rejection Sampling 13

11 Markov Chain Monte Carlo (MCMC) **13**

11.1 Metropolis 13

11.2 Metropolis-Hastings 13

11.3 Metropolis-Adjusted Langevin Algorithm (MALA) 13

11.4 Hamiltonian Monte Carlo 13

1 Introduction

1.1 Philosophy

2 Preliminary

2.1 Kronecker Product

2.2 Gradients

2.3 Jacobian

2.4 Hessian

2.5 Newton Optimization

3 Distributions

3.1 Distribution

3.2 Arcsine

3.3 Asymmetric Laplace

3.4 Bernoulli

3.5 Beta

3.6 Categorical

3.7 Cauchy

3.8 Chi Square

3.9 Conditional Model

3.10 Convolution

3.11 Data

3.12 Dirac Delta

3.13 Dirichlet

3.14 Exponential

3.15 Fisher-Snedcor (F-Distribution)

3.16 Gamma

3.17 Generator

3.18 Gumbel Softmax

3.19 Gumbel

3.20 Half Cauchy

3.21 Half Normal

3.22 Hyperbolic Secant

3.23 Langevin

3.24 Laplace

3.25 Log Cauchy

3.26 Log Laplace

3.27 Log Normal

4.1 Mixture Model

4.2 Gumbel Mixture Model

4.3 Infinite Mixture Model

5 Transforms

Transforms are invertible functions that can be applied to a random variable to change the distribution.

5.1 Transform

5.2 Inverse Transform

5.3 Affine

- **Parameters**
  - Location $\mu \in \mathbb{R}^n$
  - Scale $\sigma > 0$
- **Forward**

$$f(x) = \mu + \sigma \cdot x \tag{1}$$

- **Inverse**

$$f^{-1}(y) = \frac{y - \mu}{\sigma} \tag{2}$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \log |\sigma| \tag{3}$$

5.4 Exp

- **Parameters**
  - None
- **Forward**

$$f(x) = e^x \tag{4}$$

- **Inverse**

$$f^{-1}(y) = \log y \tag{5}$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \tag{6}$$

5.5 Expm1

- **Parameters**
  - None
- **Forward**

$$f(x) = e^x - 1 \tag{7}$$

- **Inverse**

$$f^{-1}(y) = \log(1 + y) \tag{8}$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \tag{9}$$

5.6 Gumbel

- **Parameters**
  - Location $\mu \in \mathbb{R}^n$
  - Scale $\sigma > 0$
- **Forward**

$$f(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \tag{10}$$

- **Inverse**

$$f^{-1}(y) = \mu - \sigma \cdot \log\left(-\log\left(y\right)\right) \tag{11}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -\log\left(\frac{\sigma}{-\log(y) \cdot y}\right) \tag{12}$$

5.7 Log

- **Parameters**
  - None
- **Forward**

$$f(x) = \log x \tag{13}$$

- **Inverse**

$$f^{-1}(y) = \exp y \tag{14}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -y \tag{15}$$

5.8 Logit

- **Parameters**
  - None
- **Forward**

$$f(x) = \log\left(\frac{x}{1-x}\right) \tag{16}$$

- **Inverse**

$$f^{-1}(y) = \frac{1}{1+e^{-y}} \tag{17}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = \log\left(1+e^{-y}\right) + \log\left(1+e^{y}\right) \tag{18}$$

5.9 Planar

5.10 Power

- **Parameters**
  - Power p
- **Forward**

$$f(x) = \begin{cases} e^x & p = 0 \\ (1 + x \cdot p)^{1/p} & \text{otherwise} \end{cases} \tag{19}$$

- **Inverse**

$$f^{-1}(y) = \begin{cases} \log y & p = 0 \\ y^{p-1}/p & \text{otherwise} \end{cases} \tag{20}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = \begin{cases} x & p = 0 \\ \left(\frac{1}{p} - 1\right) \cdot \log\left(x \cdot p + 1\right) & \text{otherwise} \end{cases} \tag{21}$$

5.11 Reciprocal

- **Parameters**
  - None
- **Forward**

$$f(x) = 1/x \tag{22}$$

- **Inverse**

$$f^{-1}(y) = 1/y \tag{23}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -2 \cdot \log|x| \tag{24}$$

5.12 Sigmoid

- **Parameters**
  - None
- **Forward**

$$f(x) = \frac{1}{1 + e^{-x}} \tag{25}$$

- **Inverse**

$$f^{-1}(y) = \log\left(\frac{y}{1 - y}\right) \tag{26}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -\log\left(1 + e^{-x}\right) - \log\left(1 + e^{x}\right) \tag{27}$$

5.13 SinhArcsinh

5.14 Softplus

5.15 Softsign

5.16 Square

5.17 Tanh

6 Criterion and Divergences

The criterion and divergences listed here can be used to quantify the "distance" between two distributions. Hence, in conjunction with torch optimizers, one can minimize said difference to learn the paramters of a distribution. For sake of notation clarity, p is the true distribution and q is the learned distribution. Hence we "fit" q to match p. In addition, we provide the Monte Carlo approximation.

6.1 Cross-Entropy

$$\begin{aligned} H(p, q) &= -\int p(x) \log q(x) dx \\ &= -\frac{1}{n} \sum_{x \sim p} \log q(x) \end{aligned} \tag{28}$$

6.2 Perplexity

$$\begin{aligned} H(p, q) &= \exp\left(-\int p(x) \log q(x) dx\right) \\ &= \exp\left(-\frac{1}{n} \sum_{x \sim p} \log q(x)\right) \end{aligned} \tag{29}$$

6.3 Forward KL Divergence

$$
\begin{aligned}
H(p,q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\
&= \frac{1}{n} \sum_{x \sim p} \log \frac{p(x)}{q(x)}
\end{aligned}
\tag{30}
$$

6.4 Reverse KL Divergence

$$
\begin{aligned}
H(p,q) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\
&= \frac{1}{n} \sum_{x \sim q} \log \frac{q(x)}{p(x)}
\end{aligned}
\tag{31}
$$

6.5 Jensen-Shannon Divergence

6.6 Earth Mover's Distance

7 ELBO

8 Adversarial Loss

Adversarial Losses are criterion functions that allow for sample-sample based training between models p and q. More formally, it hides a Discriminator model that attempts to discriminate between the real data from p and fake data generated from q.

8.1 Adversarial Loss

8.2 GAN Loss

8.3 MMGAN Loss

8.4 WGAN Loss

8.5 LSGAN Loss

8.6 Gradient Penalty

8.7 Spectral Norm

9 Models

9.1 Base Models

9.1.1 Model

9.2 Regression

9.2.1 Linear Regression (Normal)

9.2.2 L1 Regression (Laplace)

9.2.3 Ridge Regression (Normal + Normal Prior on Weights)

9.2.4 Lasso Regression (Normal + Laplace Prior on Weights)

9.3 Classification

9.3.1 Logistic Regression (Bernoulli)

9.3.2 Bayesian Logistic Regression (Bernoulli)

9.3.3 Softmax Regression (Categorical)

9.3.4 Bernoulli Naive Bayes (Bernoulli - Bernoulli)

9.3.5 Gaussian Naive Bayes (Multinomial - Gaussian)

9.3.6 Multinomial Naive Bayes (Bernoulli - Multinomial)

9.3.7 Linear Discriminant Analysis (Multinomial - Shared Covariance)

9.3.8 Quadratic Discriminant Analysis (Multinomial - Multivariate Gaussian)

9.4 Clustering

9.4.1 Gaussian Mixture Model

9.5 Unconstrained Matrix Factorization (Gaussian)

9.6 Principle Components Analysis

9.6.1 PCA

9.6.2 EM-PPCA

9.6.3 Variational PPCA

9.7 Generative Adversarial Networks

9.7.1 Generative Adversarial Networks

9.7.2 GAN Model

9.7.3 MMGAN Model

9.7.4 WGAN Model

9.7.5 LSGAN Model